Available online at www.sciencedirect.com
ScienceDirect
Infection, Genetics and Evolution 7 (2007) 731–735
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Short communication

A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels

Thierry de Meeûs [a,*], Jérôme Goudet [b]

[a] *Génétique et Evolution des Maladies Infectieuses, Unité Mixte de Recherche 2724, Institute de Recherche pour le Développement, Centre National de la Recherche Scientifique, Centre IRD, 911 Av d'Agropolis, BP 64501, 34394 Montpellier Cedex 5, France*
[b] *Department of Ecology & Evolution, Biophore Building, UNIL, CH-1015 Lausanne, Switzerland*

Received 19 March 2007; received in revised form 11 July 2007; accepted 12 July 2007
Available online 19 July 2007

Abstract

The populations of parasites and infectious agents are most of the time structured in complex hierarchy that lies beyond the classical nested design described by Wright's F-statistics (F_{IS}, F_{ST} and F_{IT}). In this note we propose a user-friendly step-by-step notice for using recent software (HierFstat) that computes and test fixation indices for any hierarchical structure. We add some tricks and tips for some special data kind (haploid, single locus), some other procedure (bootstrap over loci) and how to handle crossed factors.

Keywords: Fixation indices; Population structure; Hierarchy

1. Introduction

Population biologists, and among them those studying host populations, their pathogens and their vectors are interested in studying natural populations through molecular markers. This is particularly true for molecular epidemiologists because this represents the sole (or nearly so) way to study the populations they are interested in (e.g. De Meeûs et al., 2004). The most widely used parameters to infer population structure are the so-called F-statistics (Wright, 1951; Nagylaki, 1998) and their unbiased estimators (Weir and Cockerham, 1984). Classically, these parameters are defined for three hierarchical levels. The F_{IS} measures the identity (or homozygosity) of alleles within individuals within sub-populations as compared to Hardy–Weinberg expectations, it is thus a measure of deviation from local panmixia (random union of gametes producing zygotes). F_{ST} measures identity of individuals within sub-populations as compared to individuals from other sub-populations within the total population, or the total homozygosity due to the Wahlund effect. It is thus a measure of differentiation between sub-populations. Finally, F_{IT} is a measure of homozygosity of individuals in the total population and thus measures the deviation from Hardy–Weinberg due to local deviation from panmixia and Wahlund effect. The three indices are connected by the famous relationship: $(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$. Note that F_{ST} can be computed for haploids but of course not F_{IS} or F_{IT}. This can be analysed by many different free downloadable software (see Goudet, 2005). However, the population of pathogenic agents might not be well described with these three levels. In particular, several individuals (infra-population) of a pathogenic agent can colonise an individual host (e.g. a patient), different individual hosts may group into different villages themselves belonging to particular counties, states, countries, continent, etc.... In such cases, a global analysis requires another algorithm (and software implementing it).

Recently, Goudet (2005) developed a package for R (R Development Core Team, 2007) based on Yang's (1998) algorithm, which provides a convenient way to compute and test the significance of hierarchical F-statistics for any number of hierarchical levels, that he called HierFstat. However, the use of this package requires some knowledge of the R language. Now, many molecular epidemiologists are not very familiar with R and this could seriously limit the use of HierFstat and all the benefits that can come from a global analysis of such subdivided data (see Nébavi et al., 2006 for a good example).

* Corresponding author. Tel.: +33 467 4163 10; fax: +33 67 4162 99.
E-mail address: demeeus@mpl.ird.fr (T. de Meeûs).

1567-1348/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2007.07.005

While other softwares (Arlequin, GDA, TFPGA, reviewed in Excoffier and Heckel, 2006) offer the possibility to handle up to four hierarchical levels, HierFstat is the only program allowing for an unlimited number of levels, *F*-estimate and randomisation testing. There may also be other kind of subdividing factors such as date of sampling, sex of the host or the cohort it belongs to (age class), which are not hierarchical but crossed factors and will require special care. This is why in this note we propose a step-by-step and user-friendly tutorial to implement any kind of analysis with HierFstat, with special recommendations on data structure, a special interest to haploid data, how to handle single locus analyses, how to obtain bootstrap confidence intervals of the different *F* measured at different levels and how to handle crossed factors.

2. Data structure

For the following, the data should have the same format as the example file examplehier.txt (see the file as supplementary material available at http://gemi.mpl.ird.fr/SiteSGASS/deMeeus/ExampleFilesHierFstat.html) for three factor levels and five loci. Each column is separated by a tabulation, lev1, lev2 and lev3 represent different levels of population structure, lev1 being the most inclusive one but itself included in the total data set and lev3 the innermost one, but itself containing individuals. This means that individuals are grouped into different clusters of lev3, themselves included in different meta-clusters defined by lev2, which are themselves included in the partition defined by lev3. There are thus here two supplementary levels at each extreme of the hierarchy: the total population and the individuals (corresponding to F_{IT} and F_{IS}). Loc1, Loc2 ... Loc5 are the data obtained for five different loci. There may of course be more than five loci (actually five loci is a minimum for obtaining confidence intervals by bootstrap) and the number of hierarchical levels is not limited. The data file must be in text mode only. It is best if the labels used to define the state of each level are numbered sequentially, not repeating the labels (e.g. 1 1 1 1 1 1 1 2 2 2 2 2 2 3 3 3 3 3 3; not 2 2 2 2 1 1 1 1 3 3 3 3) in the relevant column. In the same way, a sequence like 1 1 1 2 2 2 1 1 1 2 2 2 1 1 1 2 2 2 should be avoided. Thus, a labelling like the one presented in Table 1 is ideal.

It is easier if missing data are coded as "NA" (upper cases as R differentiate it from lower cases). If "0" are to be used for missing data, the user needs to specify it when the file is read into R, using the option of the `read.table` command `na.string` = "0".

Table 1
Example of labels for factor levels

lev1	lev2	lev3
1	1	1
1	1	1
1	1	1
1	1	1
1	1	2
1	1	2
1	1	2
1	2	3
1	2	3
1	2	3
1	2	4
1	2	4
1	2	4
2	3	5
2	3	5
2	3	5
2	3	5
2	3	6
2	3	6
2	3	6
2	4	7
2	4	7

3. Estimating and testing hierarchical *F*-statistics

It is now assumed that you have downloaded and installed R in your computer (from http://www.r-project.org/) and the HierFstat package into it (from the menu "Package" click on "Install from a zip file" and browse where you copied the software). A good and gentle introduction to R can be found in Dalgaard (2002). Several tutorials and quick start guides can be found from R homepage at http://www.r-project.org/. And help for the different R commands may be obtained by typing the name of the command preceded by a question mark (e.g. `?library`). In the following, we also assume a Windows platform.

Launch R. From the R menu load HierFstat. You just need to click in the Menu "Package", to click on "Load Package" and on "HierFstat" (or type the command `library(hierfstat)`). You then need to go to the directory where the data to analyse are present. In the R Menu "File" Click on "Change Dir ..." and browse to the directory where the data file is present (or type `setwd(''mydir'')`, using / -not \- between folders, e.g. `setwd(''c:/myfolder/hierfstat/'')`.

You need now to load the data in R. We will use the data from the file examplehier.txt available at http://gemi.mpl.ird.fr/SiteSGASS/deMeeus/ExampleFilesHierFstat.html. This is done by typing the following command:

```
data<-read.table("examplehier.txt",header=TRUE)

attach(data)
```

This instructs R that your data file should be read and stored in the R object named data. The option "`header = TRUE`" means that you have named each column. Do respect capitalisation as the language behind R is case sensitive. The command `attach(data)` allows accessing directly the variable names. The file examplehier.txt is made of eight columns, the first three corresponding to the different hierarchical levels and the next five to the different loci (see `?read.table` for help).

It is convenient to define and name a data frame in R format that contains only loci (genetic) information. This is done by typing the following command: